DOCUMENT RESUME

ED 436 557 TM 030 416

AUTHOR	Lane, Ginny G.
TITLE	Show Me the Magnitude! The Consequences of Overemphasis on Null Hypothesis Significance Testing.
PUB DATE	1999-11-00
NOTE	19p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Point Clear, AL, November 16-19, 1999). Some pages contain light type.
PUB TYPE	Opinion Papers (120) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Effect Size; *Hypothesis Testing; *Research Methodology; *Statistical Significance
IDENTIFIERS	*Null Hypothesis

ABSTRACT

For years, researchers have debated the misinterpretation of the null hypothesis significance test (NHST). Many researchers overemphasize the results of the NHST and underemphasize or even omit effect size measures. This paper addresses the common mistaken perceptions regarding the NHST. Several common effect size estimates are discussed. A small data set is utilized to demonstrate how reliance on statistical significance without consulting effect size estimates can lead to erroneous conclusions. How interpretation of measures of effect size can provide the researcher with better information about the nature of results is discussed. (Contains 2 tables and 17 references.) (Author/SLD)

Running Head: NULL HYPOTHESIS SIGNIFICANCE TESTING AND EFFECT SIZE

Show Me the Magnitude! The Consequences of Overemphasis on Null Hypothesis Significance Testing

Ginny G. Lane

University of North Texas

Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, AL, November 17-19, 1999.

Abstract

For years, researchers have debated the misinterpretation of misapplication of the null hypothesis significance test (NHST). Many researchers overemphasize the results of the NHST and underemphasize or even omit effect size measures. This paper addresses the common misperceptions regarding the NHST. Several common effect size estimates are discussed. A small data set is utilized to demonstrate how reliance upon statistical significance without consulting effect size estimates can lead to erroneous conclusions. The author illustrates how interpretation of measures of effect size can provide the researcher with better information about the nature of results.

Show Me the Magnitude! The Consequences of Overemphasis on Null Hypothesis Significance Testing

The null hypothesis statistical significance test is a procedure that has dominated social science and educational research for the past 70 years (Kirk, 1996). It is a statistical procedure used to determine the likelihood of a given result assuming a true null hypothesis in the population of interest. Although surrounded by controversy for these 70 years, the null hypothesis significance test (henceforth referred to as NHST) has become the litmus test used by many researchers and publishers to judge the importance of a particular piece of research. Because there are misconceptions about what information can be derived from a NHST, researchers have been slack about providing more comprehensive statistical analyses, and publishers have been slack about demanding them. Moreover, those who read and interpret educational research often fail to look further than the NHST information provided to ascertain the impact of a study.

Although there are many ways in which the NHST has been misinterpreted and misapplied (Thompson, 1997), this paper addresses the most ubiquitous--that the NHST evaluates a study's magnitude of effect. From this common misperception stem two sins of omission: omission of information and omission of thoughtful analysis.

History of a Controversy

Almost since its inception, the NHST has been a procedure mired in controversy. Although accepted today as a unified theory, the current NHST procedure is an amalgamation of concepts from statisticians who were at war with one another (Nix & Barnett, 1998). The fundamental principle of testing a null hypothesis and using the p value to determine the strength of the statistic was developed by Sir Ronald Fisher in the

1920s. Jerzy Neyman and Karl Pearson later added the supporting concepts of Type I error, Type II error, and statistical power (Huberty, 1993). Fisher was philosophically opposed to the concept of a dichotomous yes/no decision based on statistical significance at a predetermined level, and there remained a bitter feud between the two camps until Fisher's death in 1962. Despite the animosity and the philosophical distinction between the two theories, textbooks began presenting the two views as a unified theory as early as the 1950s (Huberty, 1993). By the 1980s, the unified version of the NHST was so firmly entrenched in research protocol that over 90% of the articles in most psychology journals used the procedure to evaluate data (Nix & Barnette, 1998).

Even while gaining acceptance by journal editors and textbook publishers, the use of a predetermined alpha level as the dichotomous judgement for the "goodness" or "badness" of research results has been hotly debated. It is surprising that a procedure would become so widely accepted given the number of scholars who have argued its limitations (e.g., Carver 1978; Cohen, 1994; Daniel, 1998; Kirk, 1996; Morrison & Henkel, 1970; Thompson, 1997, 1998). In fact, one social scientist even referred to the NHST as "the most bone-headedly misguided procedure ever institutionalized in the rote training of science students" (Rozeboom, 1997, p.335). According to Thompson (1998b), there is now an emerging consensus among scholars regarding the limitations and widespread misapplications of the NHST. While there is some evidence that journal editors are beginning to see past the NHST, there still exists a bias in favor of data with a p calculated less than .05. And, while some journals encourage the reporting of effect size measures, very few actually require them (Kirk, 1996; Thompson, 1998b).

The Misinterpretation of the NHST

The widespread abuse of statistical significance testing stems from a fundamental misunderstanding of what information can be derived from the results of an NHST (Cohen, 1994; Kirk, 1996; Morrison & Henkel, 1970). The NHST is a procedure to determine the likelihood of a given result assuming the null hypothesis is true (Cohen, 1994; Kirk 1996; Thompson, 1997). It is a conditional probability that first assumes the null is true before determining the probability of the observed result. In statistics, one is usually inferring to a particular population from the sample. But in the case of the NHST, the direction of inference is from the population to the sample (Thompson, 1998b). One cannot assume the calculated p is a probability that the null is true because the null was pre-set to be zero. The p calculated speaks only to the observed data (under the condition of the null). Unfortunately, researchers have long labored under the assumption that the NHST says something about the population. Some erroneously interpret a statistically non-significant result as proof that the null hypothesis is true. Likewise, a statistically significant result can be erroneously taken as proof of the alternative hypothesis. Cohen (1994) and Thompson (1997) suggested that it is desperation that drives some to read more into the significance test than should be. No matter how desperately one wants proof of the population characteristics, nothing short of the actual population data will suffice.

Reviews of education and psychology journals by Thompson (1997) and Kirk (1996) showed that effect magnitude measures take a back seat to statistical significance in reported research. Even in cases where effect size is reported, the analysis and discussion is more heavily influenced by the NHST results. In addition, the majority of researchers limit effect size estimates to R^2 or η^2. Kirk (1996) correctly surmised that this is likely due to the fact that most commonly used statistical packages compute these measures. A comprehensive discussion of available effect size measures is given below.

Daniel (1998) finds further evidence of a misperception about the NHST in that the language is becoming blurred in the summary evaluation. The statistical term "significant" is being used to imply the overall impact of the study when it should only be appropriate in terms of the NHST (Kirk, 1996; Shaver 1993; Thompson, 1998b).

Statistical Significance ≠ Effect Size

Statistical significance in no way reflects the effect magnitude of a study. The two are separate but complementary procedures. They should not be used interchangeably although presentation of both effect size and results of statistical significance testing can provide much information to the reader of a research report. An accountant would never look at a company's balance sheet without also looking at the income statement because things can be hidden in one and found in the other. Likewise, effect magnitude measures yield information not found in the NHST.

Effect size is a function of the treatment. Statistical significance, on the other hand, is a function of sample size because the statistic used to determine p calculated is mathematically tied to n (sample size). Consider the computation of the t statistic: The difference between means is divided by the standard error which is computed by dividing by the square root of n. A larger n results in a smaller standard error which in turn produces a smaller divisor. A smaller divisor produces a larger t. Likewise with the F statistic: the MS between is divided by the MS within. The MS within is computed using a ratio of the error sums of squares to the error degrees of freedom, the latter of which is n - (df within – 1).. Again, following the same chain of computation, a larger n produces smaller MS within. A smaller MS within produces a smaller divisor for the F statistic ratio. Thus, even in the case of a trivial effect size, a large sample will ensure statistically significant results. If a new treatment yields a difference in means of one point then who cares if it is statistically significant?

The following examples illustrate the impact of sample size on statistical significance and the misinterpretations that can follow when thinking stops at the NHST. The data for examples 1, 2 and 3 are drawn from a hypothetical experiment involving two levels of English language instruction for three ethnic categories of immigrant students with limited English. The dependent variable is a test of verbal communication (comprehension and speaking). The data were analyzed using a two-way ANOVA with an alpha level of .05. Each experiment involves the same conditions except for sample size (n=20,30, and 190).

The results of each experiment are given in Table 1. The first line shows the results when the sample size was 20. With a sample size of 20 (10 per group) there is a mean difference of approximately 5 points. The estimated effect size .208, which is noteworthy. The null is NOT rejected because the p calculated is .076, which is larger than the preset alpha level of .05.

The second line shows results for the same type of experiment but with a sample size of 30 (just five more pupils per group). The difference in means is only four points. The effect size drops to .18, but the null hypothesis in this scenario WOULD be rejected because the p calculated is .04, which is below the alpha (.05) criterion. Although these two studies would technically support one another, the first example would not be considered for publication by many editors who are biased against statistically non-significant results.

Results of the third hypothetical experiment (n=190) show what can happen when trivial differences are found in large sample sizes. For this experiment, there is a mean difference of only one point, the effect size is near zero, yet the results are statistically significant (p calculated =04.) The null would be rejected.

Which scenario yields the most important results? It is up to the researcher who has collected the data and observed the phenomena to make this determination in light of other research in the field of immigrant education and language acquisition; nevertheless, it is reasonable to assume that a statistical effect close to zero would not be regarded as important despite the level of statistical significance of the result. Furthermore, the results must be evaluated in the context of the entire study. Assume for a moment that the treatment in example one had been a four-week course. A difference of five points (and an effect size of .20) may represent a phenomenal change in such a short time. English is critical to school success so it may be worth the risk of a Type I error to chance an improvement in such a short time. Suppose, however, the treatment had lasted a year. A five-point difference in means may not be considered substantial over such a long period, especially if the treatment is costly.

This is not to suggest that one must not be conservative when generalizing from small sample sizes. It merely suggests that rigid adherence by the research community to the $p < .05$ rule of statistical significance may discriminate against important small sample results that could very well support the findings of those fortunate enough to have larger samples or else could open doors for further research on a worthy treatment. This is especially important in those areas where large sample sizes are difficult to find (e.g., special education).

Consequences

One of the main consequences of an overreliance on the results of an NHST and an under emphasis on effect size is that good research often does not get reported. Well-designed and executed studies with appreciable effects are doomed to the reject pile merely because they do not meet the $p < .05$ rule of statistical significance that has been established as the rule of thumb in educational research (Daniel, 1998). Many (e.g., Carver, 1993; Morrison & Henkel, 1970; Nix & Barnette, 1998; Shaver; 1993) have suggested that this bias towards statistically non-significant results impedes scientific inquiry because data that could support other findings and offer some evidence of replication are not reported. Moreover, even in those instances in which results are statistically significant and studies are published, it is still too often the case that authors provide an inadequate amount of information to enable one to determine the effect size (e.g., reporting of ANOVA F statistics in absence of eta-squared values and/or sum of squares partitions to establish eta-squared). Furthermore, Thompson (1997) pointed out that routine effect size reporting will make it easier to more accurately synthesize findings via meta-analysis.

A second and more egregious consequence is that an overreliance on the NHST stunts thinking. Because naïve researchers assume the results of an NHST describe the population and evaluate the overall impact of the study, they too often stop there. Even when effect size is reported, it is limited to the two most common procedures (R^2 or η^2) that are included in statistical computer software. No other tools are considered that might remove the positive bias that exists in these two measures (Kirk, 1996). Researchers should further analyze the previous work in their field to determine the average effect size for that particular treatment or study (Cohen, 1994). The results could be evaluated in light of the expected versus the observed effect. Furthermore, as demonstrated above, researchers should make a determination based on the entire context of the study.

Suggestions: More Information and Less Rigidity

There are two dimensions to the solution as presented in the current literature (e.g., Carver, 1993; Cohen, 1980, 1994; Daniel, 1998; Shaver, 1993; Thompson, 1997). One dimension involves issues related to actual reporting of statistical analyses conducted by the researcher. The second dimension involves a paradigm shift within the publishing world and the research community at large.

Reforms Relating to Reporting of Results

The responsibility of the researcher is to go beyond the results of the NHST and provide a more comprehensive analysis of the results presented. It has been suggested by many (e.g., Carver, 1993; Cohen, 1994; Daniel, 1998; Nix & Barnette, 1998; Snyder & Lawson, 1993; Thompson, 1997) that researchers include effect magnitude estimates in their reported analysis. This would force researchers to go beyond the NHST in evaluating their results and would also afford the readers sufficient information to interpret the results in their own context.

There are many tools available to researchers to estimate the magnitude of effect of their study. Table 2 lists some of the available procedures by category. There are two categories of effect magnitude measures (a) measures of standardized effect size and (b) measures of strength of association (Kirk, 1996; Nix & Barnette, 1998; Snyder & Lawson, 1993). Measures of standardized effect size (also referred to as standardized differences) directly involve the differences between means. Measures of strength of association (also called variances accounted for) concern proportions of variance in the dependent variable associated with the independent variable. Snyder and Lawson (1993) caution that some of the more popular effect size measures (e.g., R^2 and η^2) are positively biased . These procedures tend to overestimate the population parameters. Alternatives are the unbiased measures (such as omega^2 and epsilon$^{2)}$ or correction formulas such as the Wherry, Lord, and Herzberg formulas.

In addition to reporting effect size measures, it has been suggested that confidence intervals be used to supplement the NHST in reporting research results. Kirk (1996) and Thompson (1997) pointed out that the confidence interval requires no more effort than the NHST but provides a range of values within which the true parameters are bound to lie. Hence, the confidence interval can give the researcher and the reader a reminder that there is a range of error for the results. Thompson further pointed out that, unlike p values, confidence intervals are reported in the same metric as the data and are more easily interpreted.

It has also been suggested (Daniel, 1998) that the language used in the interpretation of analysis be more precise. Even if the researcher does not intend to imply importance, ambiguous language can mislead those who read and interpret published research. It is an ironic nuisance that the term "significance" connotes importance in non-scientific English. If my checkbook is significantly out of balance I am in trouble. If I received a significant raise there would be cause for celebration. For this reason, Daniel (1998) suggested that authors insert the word "statistically" before significant when speaking in terms of study results. Other language should be used to evaluate the overall impact of the results in the particular study context.

Reforms Related to a Paradigm Shift

The second part of the solution is more complicated than the suggested reporting reforms. The research community has clung for life to the NHST. It has been the cornerstone of editorial policy for the last two decades. Something so firmly entrenched becomes habit. It is not easy to change the establishment (something that must have crossed the mind of Copernicus while he languished in prison!) The longer a practice remains, the more credibility it garners. As Frick (1996, p.379) noted in a defense of the NHST, "A way of thinking that has survived decades of ferocious attacks is likely to have some value."

Thompson (1998b) however, has found evidence of a slight shift in attitude. In 1994, APA editorial policies encouraged authors to provide measures of effect magnitude for every reported p value. In 1996 the APA appointed a task force to research the issue and make policy recommendations to foster more informed and thoughtful analyses. Kirk (1996) found at least three journals that currently require effect magnitude measures: The Journal of Experimental Education, Educational and Psychological Measurement, and The Journal of Applied Research. Daniel (1998) and Thompson (1998a) noted several other journals that have adopted such policies.

However, until the publication embargo is officially lifted, graduate committees and professors on the tenure track will continue to follow the lead of the journal editors.

Likewise, until the journal gatekeepers insist on (instead of merely encourage) deeper analyses, authors will get by with the NHST because it is quick and easy.

Conclusion

Why has a limited procedure taken such firm hold and acquired super powers? Probably because researchers have an innate need to objectify investigative work. Scientists and consumers are looking for protection from human error and judgement. However, the entire process of scientific evaluation is value-laden. The formula for the NHST may be mathematically pure, but the results offer no protection from mistake and bias.

Research in the social sciences is based on human behavior, and no matter how badly scientists need to explain human response, there will never be a foolproof way to do it. Teachers know this. Classes from year to year are never the same. What worked in 1995 may very well flop in 1999. Thinking and learning are such highly individualized processes that teachers need a vast array of methods in their pedagogical arsenal (Jensen, 1998). If I move to California next week, the language teaching method described in the experiment on page six may suddenly become relevant. The immigrants in my new town will have a different face than where I presently live in Texas. Perhaps a treatment that constituted an effect size of .20 in Dallas will suddenly yield an effect size of .44 in new surroundings.

Those who suffer under the delusion of objectivity forget the entire context of investigative research is value-laden. The questions that are asked, the measurement instruments, the study design, and the funding are all issues affecting the course of research. These issues are all related the socio-cultural context of the moment and researchers' own personal beliefs. It is ironic, then, that the researcher controls numerous factors that affect research outcomes but is asked to divorce himself or herself from the evaluation of research importance in light of an "objective" NHST. As Kirk wrote in 1996 (p.755):

> It is a curious anomaly that researchers are trusted to make a variety of complex decisions in the design and execution of an experiment, but in the name of objectivity, they are not expected or even encouraged to decide whether the data are practically significant.

References

Carver, R.P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R.P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49, 997-1003.

Daniel, L.G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. Research in the Schools, 5(2), 23-32.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.

Huberty, C.J. (1993.) Historical origins of statistical significance testing practices: The treatment of Fisher versus Neyman –Pearson views in textbooks. Journal of Experimental Education, 61, 377-333.

Jensen, E. (1998). Teaching with the brain in mind. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Morrison, D.E., & Henkel, R.E. (Eds.). (1970). The significance test controversy: A reader. Chicago: Alkine Publishing Company.

Nix, T.W., & Barnette, J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. Research in the Schools, 5(2), 3-14.

Rozeboom, W.W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), What if there were no significance tests? Mahwah, NJ: Erlbaum.

Shaver, J.P. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61, 293-316.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 113-349.

Thompson, B. (1997). If statistical significance tests are broken/misused, what practices should supplement or replace them? (Report No. TM-027-613). East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED413342)

Thompson, B (1998a). Statistical significance and effect size reporting: Portrait of a possible future. Research in the Schools, 5(2), 33-38.

Thompson, B. (1998b). Why "encouraging" effect size reporting isn't working: The etiology of researcher resistance to changing practices. (Report No. TM-026-066). East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED416214)

Table 1

Impact of Sample Size on Statistical Significance

	Sample Size	Difference in Means	Eta Squared	P Calculated	Decision
Example 1	20	5 points	.208	.076	NOT REJECT ($p>.05$)
Example 2	30	4 points	.180	.031	REJECT ($p<.05$)
Example 3	190	1 point	.022	.04	REJECT ($p<.05$)

Table 2

Procedures to Measure Magnitude of Effect *

Measures of Strength of Association (or variance accounted for)	Measures of Effect Size (or standardized differences)
r, r_{pb}: *Biased Estimate*	Cohen's d: *for T test*
R, R^2 : *Biased Estimate*	Cohen's f: *for ANOVA, ANCOVA*
η, η^2: *Biased Estimate*	Cohen's q: *for Correlation*
$\eta_{\text{multi}, \Phi}$	Cohen's h: *for Proportions*
Cohen's f^2	Cohen's w: *for Chi Square*
Contingency Coefficient	Glass's g′
Cramer's V	Hedge's g
Fisher's Z	Rosenthal and Rubin's Π
Hay's φ^2: *Unbiased Estimate*	Tang's ϕ
Kelley's Σ^2: *Unbiased Estimate*	
Kendell's W	
Lord: *Correction Formula*	
Wherry: *Correction Formula*	
Herzberg: *Correction Formula*	

*from Kirk (1996); Snyder & Lawson (1993)

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM030416

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Show Me the Magnitude! The Consequences of Overemphasis on Null Hypothesis Significance Testing	
Author(s): Ginny G Lane	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Ginny G Lane	Printed Name/Position/Title: Ginny G Lane	
Organization/Address: 2806 Yorkshire Lane, Southlake TX, 76092	Telephone: (817) 431-5137	FAX:
	E-Mail Address: ginny.lane@worldnet.att.net	Date: 11-17-99

(over)